On the ERM Principle with Networked Data

Yuanhong Wang,1,2 Yuyi Wang,3 Xingwu Liu,4,5 Juhua Pu1,2∗

1 State Key Laboratory of Software Development Environment, Beihang University, Beijing, China
2 Research Institute of Beihang University in Shenzhen, Shenzhen, China
3 Disco Group, ETH Zurich, Switzerland
4 Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
5 University of Chinese Academy of Sciences, Beijing, China

Abstract

Networked data, in which every training example involves two objects and may share some common objects with others, is used in many machine learning tasks such as learning to rank and link prediction. A challenge of learning from networked examples is that target values are not known for some pairs of objects. In this case, neither the classical i.i.d. assumption nor techniques based on complete U-statistics can be used. Most existing theoretical results of this problem only deal with the classical empirical risk minimization (ERM) principle that always weights every example equally, but this strategy leads to unsatisfactory bounds. We consider general weighted ERM and show new universal risk bounds for this problem. These new bounds naturally define an optimization problem which leads to appropriate weights for networked examples. Though this optimization problem is not convex in general, we devise a new fully polynomial-time approximation scheme (FPTAS) to solve it.

1 Introduction

“No man is an island, entire of itself ...”, the beginning of a well-known poem by the 17th century English poet John Donne, might be able to explain why social networking websites are so popular. These social media not only make communications convenient and enrich our lives but also bring us data, of an unimaginable amount, that is intrinsically networked. Social network data nowadays is widely used in research on social science, network dynamics, and as an inevitable fate, data mining and machine learning (Scott 2017). Similar examples of networked data such as traffic networks (Min and Wynter 2011), chemical interaction networks (Szklarczyk et al. 2014), citation networks (Dawson et al. 2014) also abound throughout the machine learning world.

Admittedly, many efforts have been made to design practical algorithms for learning from networked data, e.g., (Liben-Nowell and Kleinberg 2007, Macskassy and Provost 2007, Li et al. 2016, Garcia-Duran et al. 2016). However, not many theoretical guarantees of these methods have been established, which is the main concern of this paper. More specifically, this paper deals with risk bounds of classifiers trained with networked data (CLANET) whose goal is to train a classifier with examples in a data graph G. Every vertex of G is an object and described by a feature vector X ∈ X that is drawn independently and identically (i.i.d.) from an unknown distribution, while every edge corresponds to a training example whose input is a pair of feature vectors (X, X′) of the two ends of this edge and whose target value Y is in {0, 1}.

A widely used principle to select a proper model from a hypothesis set is Empirical Risk Minimization (ERM). Papa, Bellet, and Clémenc¸on (2016) establish risk bounds for ERM on complete data graphs, and the bounds are independent of the distribution of the data. These bounds are of the order O(1/n) where n is the number of vertices in the complete graph. However, in practice it is very likely that one cannot collect examples for all pairs of vertices and then G is usually incomplete, thus techniques based on complete U-processes in (Papa, Bellet, and Clémenc¸on 2016) cannot be applied and the risk bounds of the order O(1/n) are no longer valid in this setting. By generalizing the moment inequality for U-processes to the case of incomplete graphs, we prove novel risk bounds for the incomplete graph.

Usually, every training example is equally weighted (or unweighted) in ERM, which seems much less persuasive when the examples are networked, in particular when the graph is incomplete. But, most existing theoretical results of learning from networked examples are based on the unweighted ERM (Usunier, Amini, and Gallinari 2006, Ralaivola, Szafranski, and Stempfel 2009), and their bounds are of the order O(1/√ν∗(DG)/m) where DG is the line graph of G and ν∗ is the fractional chromatic number of DG (see Section A in the online appendix1) and m is the number of training examples. In order to improve this bound, Wang, Guo, and Ramon (2017) propose weighted ERM which adds weights to training examples according to the data graph, and show that the risk bound for weighted ERM can be of the order O(1/√ν∗(G)) where ν∗(G) is the fractional matching number of G, so using weighted ERM networked data can be more effectively exploited than the equal weighting method, as basic graph theory tells us ν∗(DG) ≥ m/χ∗(DG). However, Wang, Guo, and Ramon (2017) (in fact, Usunier, Amini, and Gallinari (2006) and Ralaivola, Szafranski, and

∗Corresponding author
Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
Stempfel (2009) also assume that any two examples can be arbitrarily correlated if they share a vertex, which cannot lead to an $O(\log(n)/n)$ bound when the graph is complete.

We show that the “low-noise” condition, also called the Mammen-Tsybakov noise condition (Mammen and Tsybakov 1998), which is commonly assumed in many typical learning problems with networked data, e.g., ranking (Clémençon, Lugosi, and Vayatis 2008) and graph reconstruction (Papa, Bellet, and Clémençon 2016), can be used to reasonably bound the dependencies of two examples that share a common vertex and then leads to tighter risk bounds.

In summary, in this paper we mainly

- prove new universal risk bounds for CLANET which
  - can be applied to learning from networked data even if the data graph is incomplete;
  - exploit the property of the “low-noise” condition, and then become tighter than previous results;
  - allow non-identical weights on different examples, so it is possible to achieve better learning guarantee by choosing these weights.
- formulate a non-convex optimization problem inspired by our new risk bounds (because our risk bounds depend on the weights added to every training example, and a better weighting scheme leads to a tighter bound), and then we also design a new efficient algorithm to obtain an approximate optimal weighting vector and show that this algorithm is a fully polynomial-time approximation scheme for this non-convex program.

## 2 Intuitions

We now have a look at previous works that are closely related to our work, as shown in Table 1, and present the merits of our method. Biau and Bleakley (2006), Clémençon, Lugosi, and Vayatis (2008) and Papa, Bellet, and Clémençon (2016) deal with the case when the graph is complete, i.e., the target value of every pair of vertices is known. In this case, Clémençon, Lugosi, and Vayatis (2008) formulate the “low-noise” condition for the ranking problem and demonstrate that this condition can lead to tighter risk bounds by the moment inequality for $U$-processes. Papa, Bellet, and Clémençon (2016) further consider the graph reconstruction problem introduced by Biau and Bleakley (2006) and show this problem always satisfies the “low-noise” condition.

If the graph is incomplete, one can use either Janson’s decomposition (Janson 2004, Usunier, Amini, and Gallinari 2006, Ralaivola, Szafranski, and Stempfel 2009, Ralaivola and Amini 2015) or the fractional matching approach by Wang, Guo, and Ramon (2017), if several examples share a vertex, then we are likely to put less weight on them because the influence of this vertex to the empirical risk should be bounded. Otherwise, if we treat every example equally, then these dependent examples may dominate the training process and lead to the risk bounds that do not converge to 0 (see the example in Section 6.2).

Uniform Bounds Ralaivola and Amini (2015) prove an entropy-base concentration inequality for networked data using Janson’s decomposition, but the assumption there is usually too restrictive to be satisfied (see Section A.2 in the online appendix). To circumvent this problem, our method uses the “low-noise” condition (also used in (Papa, Bellet, and Clémençon 2016)) to establish uniform bounds, in absence of any restrictive condition imposed on the data distribution.

## 3 Preliminaries

In this section, we begin with the detailed probabilistic framework for CLANET, and then give the definition of weighted ERM on networked examples.

### 3.1 Problem Statement

Consider a graph $G = (V, E)$ with a vertex set $V = \{1, \ldots, n\}$ and a set of edges $E \subseteq \{(i,j) : 1 \leq i \neq j \leq n\}$. For each $i \in V$, a continuous random variable (r.v.) $X_i$, taking its values in a measurable space $\mathcal{X}$, describes features of vertex $i$. The $X_i$’s are i.i.d. r.v.’s following some unknown distribution $P_\mathcal{X}$. Each pair of vertices $(i, j) \in E$ corresponds to a networked example whose input is a pair $(X_i, X_j)$ and target value is $Y_{i,j} \in \mathcal{Y}$. We focus on binary classification in this paper, i.e., $\mathcal{Y} = \{0, 1\}$. Moreover, the distribution of
target values only depends on the features of the vertices it contains but does not depend on features of other vertices, that is, there is a probability distribution \( P_{Y|X} \) such that for every pair \((i, j) \in E\), the conditional probability

\[
P[Y_{ij} = y \mid x_1, \ldots, x_n] = P_{Y|X} [y, x_i, x_j].
\]

**Example 1** (pairwise ranking). Liu (2009) categorize ranking problems into three groups by their input representations and loss functions. One of these categories is pairwise ranking that learns a binary classifier telling which document is better in a given pair of documents. A document can be described by a feature vector from the \( X \) describing title, volume, … The target value (rank) between two documents, that only depends on features of these two documents, is 1 if the first document is considered better than the second, and 0 otherwise.

The training set \( S := \{(X_i, X_j, Y_{ij})\}_{(i, j) \in E} \) is dependent copies of a generic random vector \((X_1, X_2, Y_{12})\) whose distribution \( P = P_X \otimes P_X \otimes P_{Y|X^2} \) is fully determined by the pair \((P_X, P_{Y|X^2})\). Let \( \mathcal{R} \) be the set of all measurable functions from \( X^2 \) to \( Y \) and for all \( r \in \mathcal{R} \), the loss function \( \ell(r, (x_1, x_2, y_{12})) = \mathbb{1}_{y_{12} \neq r(x_1, x_2)} \).

Given a graph \( G \) with training examples \( S \) and a hypothesis set \( R \subseteq \mathcal{R} \), the CLANET problem is to find a function \( r \in R \), with risk

\[
L(r) := \mathbb{E} [\ell(r, (X_1, X_2, Y_{12}))]
\]

that achieves a comparable performance to the Bayes rule

\[
r^* = \text{arg inf}_{r \in \mathcal{R}} L(r) = \mathbb{1}_{\eta(x_1, x_2) \geq 1/2},
\]

whose risk is denoted by \( L^* \), where \( \eta(x_1, x_2) = P_{Y|X^2}[1, x_1, x_2] \) is the regression function.

The main purpose of this paper is to devise a principled way to select a classifier \( r \) from the hypothesis set \( R \) and establish bounds for its excess risk \( L(r) - L^* \).

**Definition 1** ("low-noise" condition). Let us consider a learning problem, in which the hypothesis set is \( \mathcal{F} \) and the Bayes rule is \( f^* \). With slightly abusing the notation, this problem satisfies the "low-noise" condition if \( \forall f \in \mathcal{F}, L(f) - L^* \geq C \| f - f^* \|_q^q \) where \( C \) is a positive constant.

As mentioned, the "low-noise" condition can lead to tighter risk bounds. For this problem, we show that the "low-noise" condition for the i.i.d. part of the Hoeffding decomposition (Hoeffding 1948) of its excess risk can be always obtained if the problem is symmetric (see Lemma 2).

**Definition 2** (symmetry). A learning problem is symmetric if for every \( x_i, x_j \in \mathcal{X}, y_{ij} \in \mathcal{Y} \) and \( r \in \mathcal{R} \), \( \ell(r, (x_i, x_j, y_{ij})) = \ell(r, (x_j, x_i, y_{ij})) \).

Many typical learning problems are symmetric. For example, pairwise ranking problem with symmetric functions \( r \) in the sense that \( r(X_1, X_2) = 1 - r(X_2, X_1) \) satisfies the symmetric condition.

### 3.2 Weighted ERM

ERM aims to find the function from a hypothesis set that minimizes the empirical estimator of (1) on the training examples \( S = \{(X_i, X_j, Y_{ij})\}_{(i, j) \in E} \):

\[
L_m(r) := \frac{1}{m} \sum_{(i, j) \in E} \ell(r, (X_i, X_j, Y_{ij})).
\]

where \( m \) is the number of training examples. In this paper, we consider its weighted version, in which we put weights on the examples and select the minimizer \( r_w \) of the weighted empirical risk

\[
L_w(r) := \frac{1}{\| w \|_1} \sum_{(i, j) \in E} w_{ij} \ell(r, (X_i, X_j, Y_{ij})).
\]

where \( w \) is a fractional matching of \( G \) and \( \| w \|_1 > 0 \).

**Definition 3** (fractional matching). Given a graph \( G = (V, E) \), a fractional matching \( w \) is a non-negative vector \((w_{ij})_{(i, j) \in E}\) that for every vertex \( i \in V \), \( \sum_{j, (i, j) \in E} w_{ij} \leq 1 \).

### 4 Universal Risk Bounds

In this section, we use covering numbers as the complexity measurement of hypothesis sets to prove that tighter universal risk bounds are always attained by the minimizers of the weighted empirical risk (3).
4.1 Covering Numbers

The excess risk $L(r_w) - L^*$ depends on the hypothesis set $R$ whose complexity can be measured by covering number (Cucker and Zhou 2007). A similar but looser result using VC-dimension (Vapnik and Chervonenkis 1971) can be obtained as well.

Definition 4 (covering numbers). Let $(F, L_p)$ be a metric space with $L_p$-pseudometric. We define the covering number $N(F, L_p, \epsilon)$ be the minimal $n \in \mathbb{N}$ such that there exist $n$ disks in $F$ with radius $\epsilon$ covering $F$. If the context is clear, we simply denote $N(F, L_p, \epsilon)$ by $N_p(F, \epsilon)$.

In this paper, we focus on the $L_\infty$ covering number $N_\infty(F, \epsilon)$ and suppose that it satisfies the following assumption.

Assumption 1. There exists a nonnegative number $\beta < 1$ and a constant $K$ such that $\log N_\infty(F, \epsilon) \leq K \epsilon^{-\beta}$ for all $\epsilon \in (0, 1]$.

Similar to (Massart and Nédélec 2006) and (Rejchel 2012), we restrict to $\beta < 1$, whereas in the empirical process theory this exponent usually belongs to $(0, 2)$. This restriction is needed to prove Lemma 1, which involves the integral of $\log N_\infty(F, \epsilon)$ through 0. Dudley (1974), Korostelev and Tsybakov (1993) and Mammen and Tsybakov (1995) presented various examples of classes $F$ satisfying Assumption 1. We also refer interested readers to (Mammen and Tsybakov 1998, p. 1813) for more concrete examples of hypothesis classes with smooth boundaries satisfying Assumption 1.

4.2 Risk Bounds

Now we are ready to show the tighter risk bounds for weighted empirical risk by the following theorem.

Theorem 1 (risk bounds). Let $r_w$ be a minimizer of the weighted empirical risk $L_w$ over a class $R$ that satisfies Assumption 1. There exists a constant $C > 0$ such that for all $\delta \in (0, 1]$, with probability at least $1 - \delta$, the excess risk of $r_w$ satisfies

$$L(r_w) - L^* \leq 2 \sqrt{\log(1/\delta)} \left( \frac{K' C \log(1/\delta)}{\epsilon} \right)^{1/(1+\beta)} \left( \frac{\|w\|_{\infty}}{\epsilon} \right)^{1/(1+\beta)} + \frac{\log(1/\delta)}{\epsilon},$$

where $\|w\|_{\infty} = \max_i \sqrt{\sum_{j \in E} w_{i,j}^2}$ and $K' = \max(K, \sqrt{K}, K^{1/(1+\beta)})$.

According to Theorem 1, if the parameter $\delta$ is greater than the value $\exp \left( - \min \left( \frac{\|w\|_2}{\|w\|_{\infty}}, \frac{\|w\|_2}{\|w\|_{\infty}}, \frac{\|w\|_2}{\|w\|_{\infty}} \right) \right)$, then the risk bounds above are of the order $O \left( \left( \frac{\|w\|_2}{\|w\|_{\infty}} \right)^{1/(1+\beta)} + \frac{\log(1/\delta)}{\epsilon} \right)$. In this case, our bounds are tighter than $O(1/\sqrt{\|w\|_1})$ as $\|w\|_2/\|w\|_{\infty} \leq 1/\sqrt{\|w\|_1}$ (recall that $w$ must be a fractional matching and $0 < \beta < 1$). If $G$ is complete and every example is equally weighted, the bounds of the order $O \left( \left( \frac{\|w\|_2}{\|w\|_{\infty}} \right)^{1/(1+\beta)} \right)$ achieve the same results as in (Papa, Bellet, and Clémenc`on 2016)\(^2\)

Remark. Theorem 1 provides universal risk bounds no matter what the distribution of the data is. The factor of 2 in front of the approximation error $\inf_{r \in R} L(r) - L^*$ has no special meaning and can be replaced by any constant larger than 1 with a cost of increasing the constant $C$. Wang, Guo, and Ramon (2017) obtain risk bounds that has a factor 1 in front of the approximation error part, but in their result the bound is $O(1/\sqrt{\|w\|_1})$. Hence, Theorem 1 improves their results if the approximation error does not dominate the terms in the bounds.

In the rest of this section, we outline the main ideas to obtain this result. We first define $q_r(x_1, x_2, y_1, 2) := \ell(r, x_1, x_2, y_1, 2) - \ell(r^*, x_1, x_2, y_1, 2)$ for every $(x_1, x_2, y_1, 2) \in \mathcal{X} \times \mathcal{X} \times \mathcal{Y}$ and let $\Lambda(r) := L(r) - L^* = \mathbb{E}[q_r(X_1, X_2, Y_1, 2)]$ be the excess risk with respect to the Bayes rule. Its empirical estimate by weighted ERM is

$$\Lambda_w(r) = L_w(r) - L_w(r^*) = \frac{1}{\|w\|_1} \sum_{i,j \in E} w_{i,j} q_r(X_i, X_j, Y_i, j).$$

By Hoeffding’s decomposition (Hoeffding 1948), for all $r \in R$, one can write

$$\Lambda_w(r) = T_w(r) + U_w(r) + \tilde{U}_w(r),$$

where

$$T_w(r) = \Lambda(r) + \frac{2}{\|w\|_1} \sum_{i,j \in E} w_{i,j} h_r(X_i, X_j)$$

is a weighted average of i.i.d. random variables with $h_r(X_i) = \mathbb{E}[q_r(X_i, X_j, Y_i, j) | X_i] - \Lambda(r)$,

$$U_w(r) = \frac{1}{\|w\|_1} \sum_{i,j \in E} w_{i,j} \tilde{h}_r(X_i, X_j)$$

is a weighted degenerated (i.e., the symmetric kernel $\tilde{h}_r(X_i, x_j)$ such that $\mathbb{E}[\tilde{h}_r(X_i, X_j) | X_i = x_j] = 0$ for all $x_i \in \mathcal{X}$) $U$-statistic $\tilde{h}_r(X_i, X_j) = \mathbb{E}[q_r(X_i, X_j, Y_i, j) | X_i, X_j] - \Lambda(r) - h_r(X_i) - \tilde{h}_r(X_i)$ and

$$\tilde{U}_w(r) = \frac{1}{\|w\|_1} \sum_{i,j \in E} w_{i,j} \tilde{h}_r(X_i, X_j, Y_i, j)$$

with a degenerated kernel $\tilde{h}_r(X_i, X_j, Y_i, j) = q_r(X_i, X_j, Y_i, j) - \mathbb{E}[q_r(X_i, X_j, Y_i, j) | X_i, X_j]$. In the following, we bound the three terms $T_w, U_w$ and $\tilde{U}_w$ in (5) respectively.

Lemma 1 (uniform approximation). Under the same assumptions as in Theorem 1, for any $\delta \in (0, 1/e)$, we have with probability at least $1 - \delta$,

$$\sup_{r \in R} U_w(r) \leq \frac{\max(K, \sqrt{K}) C_1}{1 - \beta} \max \left( \frac{\|w\|_2 \log(1/\delta)}{\|w\|_1} \right),$$

\(^2\)They consider the same range of $\delta$. 4247
\[ \frac{\|w\|_{\text{max}}(\log(1/\delta))^{3/2}}{\|w\|_1}, \quad \frac{\|w\|_{\infty}(\log(1/\delta))^2}{\|w\|_1} \]

and

\[ \sup_{r \in \mathcal{R}} \left( \frac{\|w\|_{\text{max}}(\log(1/\delta))^{3/2}}{\|w\|_1} + \max \left( \frac{\|w\|_{\text{max}}(\log(1/\delta))^2}{\|w\|_1}, \frac{\|w\|_{\infty}(\log(1/\delta))^2}{\|w\|_1} \right) \right) \]

where \( C_1, C_2 < +\infty \) are constants.

To prove Lemma 1, we show that \( U_w(r) \) and \( \tilde{U}_w(r) \) can be bounded by Rademacher chaos using classical symmetrization and randomization tricks combined with the decoupling method. We handle these Rademacher chaos by generalizing the moment inequality for \( U \)-statistics in (Clémencçon, Lugosi, and Vayatis 2008). Specifically, we utilize the moment inequalities from (Boucheron et al. 2005) (Clémencçon, Lugosi, and Vayatis 2008). Specifically, we utilize the moment inequalities from (Boucheron et al. 2005) (Clémencçon, Lugosi, and Vayatis 2008). Specifically, we utilize the moment inequalities from (Boucheron et al. 2005) (Clémencçon, Lugosi, and Vayatis 2008).

Let \( A \) be the event that

\[ \sup_{r \in \mathcal{R}} |U_w(r)| \leq \kappa_1, \]

where

\[ \kappa_1 = \frac{C_1}{1 - \beta} \max \left( \frac{\|w\|_2 \log(1/\delta)}{\|w\|_1}, \frac{\|w\|_{\text{max}}(\log(1/\delta))^{3/2}}{\|w\|_1}, \frac{\|w\|_{\infty}(\log(1/\delta))^2}{\|w\|_1} \right) \]

for an appropriate constant \( C_1 \). Then by Lemma 1, \( P[A] \geq 1 - \delta/4 \). Similarly, let \( B \) be the event that

\[ \sup_{r \in \mathcal{R}} |\tilde{U}_w(r)| \leq \kappa_2, \]

where

\[ \kappa_2 = \frac{C_2}{1 - \beta} \left( \frac{\|w\|_2 + \max (\|w\|_{\text{max}}(\log(1/\delta))^{3/2}, \|w\|_{\infty}(\log(1/\delta))^2)}{\|w\|_1} \right) \]

for an appropriate constant \( C_2 \). Then \( P[B] \geq 1 - \delta/4 \).

By (5), it is clear that, if both \( A \) and \( B \) happen, \( r_w \) is a \( \rho \)-minimizer of \( T_w(r) \) over \( r \in \mathcal{R} \) in the sense that the difference between the value of this latter quantity at its minimum and \( r_w \) is at most \( (\kappa_1 + \kappa_2) \). Then, from Lemma 3, with probability at least \( 1 - \delta/2 \), \( r_w \) is a \((\kappa_1 + \kappa_2)\)-minimizer of \( T_w(r) \), which the result follows.

An intuition obtained from our result is how to choose weights for networked data. By Theorem 1, to obtain tight risk bounds, we need to maximize \( \|w\|_1 \) (under the constraint that this weight vector is a fractional matching), which resembles the result of (Wang, Guo, and Ramon 2017) (but they only need to maximize \( \|w\|_1 \), and this is why they end in the \( O(1/\sqrt{\nu^*(G)}) \) bound), while making \( \|w\|_2, \|w\|_{\text{max}}, \|w\|_{\infty} \) as small as possible, which appears to suggest putting nearly average weights on examples and vertices respectively. These two objectives, maximizing \( \|w\|_1 \) and minimizing \( \|w\|_2, \|w\|_{\text{max}}, \|w\|_{\infty} \), seem to contradict each other. In the next section, we discuss how to solve this problem.

## 5 Weighting Vector Optimization

In this section, we first formulate the optimization problem that minimizes the risk bounds in Theorem 1. Although this optimization problem is not convex unless \( \beta = 0 \), which usually means that there is no general efficient way to solve it, we devise a fully polynomial-time approximation scheme (FPTAS) to solve it.

### Definition 5 (FPTAS).
An algorithm \( A \) is a FPTAS for a minimization problem \( \Pi \), if for any input \( \mathcal{I} \) of \( \Pi \), and \( \epsilon > 0 \), \( \Pi \) finds a solution \( s \) in time polynomial in both the size of \( \mathcal{I} \) and \( 1/\epsilon \) that satisfies \( f_{\Pi}(s) \leq (1 + \epsilon) \cdot f_{\Pi}(s^*) \), where \( f_{\Pi} \) is the (positive) objective function of \( \Pi \) and \( s^* \) is an optimal solution for \( \Pi \).
5.1 Optimization Problem

According to Theorem 1, given a graph $G$, $\beta \in (0, 1)$ and $\delta \in (0, 1]$, one can find a good weighting vector with tight risk bounds by solving the following program:

$$\min_w \frac{1}{\|w\|_1} \left( \|w\|_1^{\beta/(1+\beta)} + \max \left( \|w\|_2, \|w\|_\infty \right) \right)$$

subject to

$$\forall (i, j) \in E, w_{i,j} \geq 0 \quad \text{and} \quad \forall i, \sum_{j: (i,j) \in E} w_{i,j} \leq 1 \quad \text{(7)}$$

To get rid of the fraction of norms in the program above, we consider a distribution $p$ on edges $p_{i,j} := w_{i,j}/\|w\|_1$ and then $\|w\|_1 \leq 1/\max_{i=1,\ldots,n} \sum_{j:(i,j) \in E} p_{i,j}$. Every distribution $p$ corresponds to a valid weighting vector $w$. By introducing two auxiliary variables $a$ and $b$, solving the original program (7) is equivalent to solving

$$\min_{a,b,p} \quad a^{1/(1+\beta)} + b$$

subject to

$$\forall (i, j) \in E, p_{i,j} \geq 0$$

$$\forall (i, j) \in E, p_{i,j} \log(1/\delta) - b \leq 0$$

$$\forall i, \sum_{j: (i,j) \in E} p_{i,j} - a \leq 0$$

$$\forall i, \left( \sum_{j: (i,j) \in E} p_{i,j}^2 \log(1/\delta) \right)^{1/2} - b \leq 0$$

$$\|p\|_2 - b \leq 0 \quad \text{and} \quad \sum_{(i,j) \in E} p_{i,j} = 1 \quad \text{(8)}$$

Note that the constraints are all convex. If $\beta = 0$, e.g., the hypothesis set is finite, then the objective function becomes linear and thus (8) is a convex optimization program that can be solved by some convex optimization method (see e.g., Boyd and Vandenberghe 2004) such as interior-point method.

If $\beta > 0$, the objective function is not convex any more. In fact, the program (8) becomes a concave problem that may be optimized globally by some complex algorithms (Benson 1995, Hoffman 1981) that often need tremendous computation. Instead, one may only need to approximate it using some efficient methods, e.g., Concave-Convex Procedure (Yuille 2001) and Coordinate Descent (Wright 2015). However, these methods lack in complexity analysis and may lead to a local optimum.

5.2 A Fully Polynomial-time Approximation Scheme

To solve the program (8) efficiently, we propose Algorithm 1 and show that it is a fully polynomial-time approximation scheme for (8).

**Theorem 2.** Algorithm 1 is a FPTAS for the program (8).

\begin{algorithm}
\caption{FPTAS for weighting vector optimization.}
\begin{algorithmic}[1]
\State \textbf{Input:} $\epsilon, \beta, \delta$ and a graph $G$ that contains $n$ vertices and $m$ edges.
\State \textbf{Output:} An approximate optimal weighting vector $\bar{p}$ for the program (8).
\State 1: Solve the following linear program (LP) efficiently\footnote{For example, some interior-point method.}, and obtain an $\epsilon$-approximation of this optimal solution.
\State \quad \min_{a,p} a
\State \quad \text{s.t.} \quad \forall (i, j) \in E, p_{i,j} \geq 0
\State \quad \forall (i, j) \in E, p_{i,j} \log(1/\delta) - a \leq 0
\State \quad \forall i, \sum_{j:(i,j) \in E} p_{i,j} - a \leq 0
\State \quad \|p\|_2 - a \leq 0 \quad \text{and} \quad \sum_{(i,j) \in E} p_{i,j} = 1 \quad \text{(9)}
\State 2: Let $\text{Grid} := \{a_{\min} + i \cdot \epsilon(1 + \beta)/n \mid i \in \mathbb{N} \text{ and } \min(a_{\min}, b_{\max})/\epsilon(1 + \beta) \}$ and $\text{Solutions} := \emptyset$.
\State 3: for $a \in \text{Grid}$ do
\State \quad 4: Use some efficient interior-point method to obtain an $\epsilon$-approximation of the following program and add the solution $(a, b, p)$ into $\text{Solutions}$.\footnote{To find an optimal solution for this program, use an interior-point method.}
\State \quad \min_{b,p} b
\State \quad \text{s.t.} \quad \forall (i, j) \in E, p_{i,j} \geq 0
\State \quad \forall (i, j) \in E, p_{i,j} \log(1/\delta) - b \leq 0
\State \quad \forall i, \sum_{j:(i,j) \in E} p_{i,j} - b \leq 0
\State \quad \|p\|_2 - b \leq 0 \quad \text{and} \quad \sum_{(i,j) \in E} p_{i,j} = 1 \quad \text{(10)}
\State 5: return the vector $\bar{p}$ which makes $a^{1/(1+\beta)} + b$ smallest from $\text{Solutions}$.
\end{algorithmic}
\end{algorithm}

**Proof.** We first analyze the running time of this algorithm.

Note that $1/n \leq a \leq 1$ if the graph is not empty. In Algorithm 1, we first divide the problem into at most

$$\frac{1 - 1/n}{\epsilon(1 + \beta)/n} = \frac{n - 1}{\epsilon(1 + \beta)}$$

convex programs, each of which produces an $\epsilon$-approximate solution by some interior-point method. Since interior-point method is FPTAS for convex problems (Boyd and Vandenberghe 2004), solving each of these programs needs polynomial time in the problem size $m + n + 1/\epsilon$. Thus, the complexity of Algorithm 1 is also polynomial in $m + n$ and $1/\epsilon$.

Now we show that this algorithm indeed results in an $\epsilon$-approximation of this optimal solution.

For any optimal solution $(a^*, b^*, p^*)$, if $a^*$ achieves minimum for the program (9), we can find $a'$ in $\text{Grid}$ (actually
\[ (a')^{1/(1+\beta)} \leq (1 + \epsilon)^{1/(1+\beta)} (a^*)^{1/(1+\beta)} \leq (1 + \epsilon)(a^*)^{1/(1+\beta)}. \] (11)

Otherwise, we can also find \( a' \) in Grid such that \( a^* \leq a' < a^* + \epsilon(1 + \beta)/n \) and thus

\[ (a')^{1/(1+\beta)} \leq (a^*)^{1/(1+\beta)} + \epsilon(1 + \beta)/n \]

\[ \leq (1 + \epsilon)(a^*)^{1/(1+\beta)}. \] (12)

The third inequality follows from the fact that \( 1/n \leq a^* \). We assume that the optimal solution for the program (10) is \( b = b' \) when we fix \( a = a' \). Because \((a^*, b^*)\) is feasible and \( a' > a^* \), \((a', b')\) is always a feasible solution for the program (10), which leads to \( b' \leq b' \). Besides, interior-point method can produce an \( \epsilon \)-approximate solution \( b^* \) such that

\[ b'' \leq (1 + \epsilon)b' \leq (1 + \epsilon)b^*, \] (13)

Finally, we select the best approximate weighting vector \( \tilde{p} \) from all solutions in Solutions. Combining (11), (12) and (13), we have the objective value for \( \tilde{p} \)

\[ (a_\tilde{p})^{1/(1+\beta)} + b_{\tilde{p}} \leq (a^*)^{1/(1+\beta)} + b'' \]

\[ \leq (1 + \epsilon)(a^*)^{1/(1+\beta)} + b^*. \] \hfill \( \square \)

6 Discussion

In this section, we first show that, according to our bounds, equal weighting is indeed the best weighting scheme for complete graphs. Then, we discuss the performance of this equal weighting scheme when the graph is incomplete.

6.1 Complete Graphs

When graph \( G \) is complete, weighting all examples equally gives the best risk bound, as all the terms \( \max_{i=1, \ldots, n} \sum_{j:(i,j) \in E} p_{i,j} \), \( \|p\|_2 \), \( \|p\|_{\text{max}} \) and \( \|p\|_{\infty} \) achieve minimum. Compared to the results in (Wang, Guo, and Ramon 2017), our theory puts additional constrains on \( \|p\|_2 \), \( \|p\|_{\text{max}} \) and \( \|p\|_{\infty} \) which encourages weighting examples fairly in this case, as illustrated in Figure 1. Besides, this scheme, which coincides with \( U \)-statistics that average the basic estimator applied to all sub-samples, produces the smallest variance among all unbiased estimators (Hoeffding 1948).

6.2 Equal Weighting

Let us discuss further the equal weighting scheme that gives every example the same weight. Denote by \( \Delta(G) \) the maximum degree of \( G \) (note that this is not the maximum degree of \( D_G \)) and let \( p_{i,j} = 1/m \) (recall that \( m \) is the number of examples) for all \((i, j) \in E\). According to program (8), using equal weighting scheme, the risk bounds are of the order

\[ O \left( \left( \frac{\Delta(G)}{m} \right)^{1/(1+\beta)} + \frac{1}{\sqrt{m}} \right), \] (14)

if \( \delta \in (\exp(-m/\Delta(G)), 1] \). In some cases, such as bounded degree graphs and complete graphs, this scheme provides reasonable risk bounds. Note that \( \Delta(G) \) is smaller than the maximum size of cliques in its corresponding line graph \( D_G \) and \( \chi^*(D_G) \) is larger than the maximum size of cliques in \( D_G \); these bounds above are always better than the bounds of the order \( O(\sqrt{\Delta(G)/m}) \) built by Janson’s decomposition.

However, as argued in Section 2, one can construct examples to illustrate that if we use the equal weighting strategy when \( \Delta(G) \) is large (e.g., if it is linear to \( m \)), the risk bounds (14) are very large and do not converge to 0, while this problem can be solved by simply using a better weighting strategy.

Example 2. Consider a data graph with \( |E| = m \gg 1 \) and \( E \) consists of \( m/2 \) disjoint edges and \( m/2 \) edges sharing a common vertex, then \( \Delta(G) = m/2 \). Using the equal weighting scheme, the risk bounds are of the order \( O(1) \) that is meaningless. A much better weighting scheme of this case is to weight the examples of disjoint edges with \( 2/(m+2) \) while weight the examples of adjacent edges with \( 4/(m(m+2)) \), which provides risk bounds of the order \( O \left( \left( \frac{\Delta(G)}{m^2} \right)^{1/(1+\beta)} + \sqrt{1/m} \right) \).

7 Conclusion

In this paper, we consider weighted ERM of the symmetric CLANET problem and establish new universal risk bounds under the “low-noise” condition. These new bounds are tighter in the case of incomplete graphs and can be degenerate to the known tightest bound when graphs are complete. Based on this result, one can train a classifier with a better risk bound by putting proper weights on training examples. We propose an efficient algorithm to obtain the approximate optimal weighting vector and prove that the algorithm is an FPTAS for the weighting vector optimization problem. Finally, we discuss two cases to show the merits of our new risk bounds.


Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful comments. This work was supported by Guangdong Shannon Intelligent Tech. co., Ltd., National Key R&D Program of China (2017YFC0803700), National Natural Science Foundation of China (61502320, 61420106013, 61521092) and Science Foundation of Shenzhen City in China (JCYJ20160419152942010).

References


